Deep learning does not outperform classical machine learning for cell-type annotation

Niklas D. Köhler*[1], Maren Büttner*[1], and Fabian J. Theis[1,2,]

1 Institute of Computational Biology, Helmholtz Center Munich for Environmental Health, Neuherberg, Germany
2 Department of Mathematics, TU Munich, Germany
* these authors contributed equally
 e-mail address: fabian.theis@helmholtz-muenchen.de

Abstract

Deep learning has revolutionized image analysis and natural language processing with remarkable accuracies in prediction tasks, such as image labeling or word identification. The origin of this revolution was arguably the deep learning approach by the Hinton lab in 2012, which halved the error rate of existing classifiers in the then 2-year-old ImageNet database[1]. In hindsight, the combination of algorithmic and hardware advances with the appearance of large and well-labeled datasets has led up to this seminal contribution.
The emergence of large amounts of data from single-cell RNA-seq and the recent global effort to chart all cell types in the Human Cell Atlas has attracted an interest in deep-learning applications. However, all current approaches are unsupervised, i.e., learning of latent spaces without using any cell labels, even though supervised learning approaches are often more powerful in feature learning and the most popular approach in the current AI revolution by far.
Here, we ask why this is the case. In particular we ask whether supervised deep learning can be used for cell annotation, *i.e.* to predict cell-type labels from single-cell gene expression profiles. After evaluating 6 classification methods across 14 datasets, we notably find that deep learning does not outperform classical machine-learning methods in the task. Thus, cell-type prediction based on gene-signature derived cell-type labels is potentially too simplistic a task for complex non-linear methods, which demands better labels of functional single-cell readouts. We, therefore, are still waiting for the “ImageNet moment” in single-cell genomics.

Main

Single-cell genomics is a success story in global scientific collaborative activities and research, combining advances in genomics, miniaturization, and data science. The widespread availability of transcriptome readouts from various cells, tissues, and species nowadays allows the integration and comparison of such data using machine-learning methods[2–6]. Successful data integration in turn facilitates reliable annotation of cell types at new datasets levels of quality as compared with those of human experts. Annotating cell types is a classical supervised learning problem, which can be addressed by using numerous classification models, such as logistic regression, support vector machines, and random forests. However, with increasing data volumes, we can soon approach data volume thresholds where deep learning models outperform classical machine-learning models in cell-type annotation tasks. Here we explore whether we have reached this threshold by comparing classical machine-learning approaches with state-of-the-art deep-learning architectures in a cell-type annotation task. In our comparison study of classifier types, we present a strategy for deriving optimal parameters for a given model to predict cell-type annotation, i.e., what level of accuracy can we expect in the best-case scenario from a given model trained using a specific dataset. We applied this testing framework to three different biological systems with increasing sample sizes: mouse embryonic development[7–11] (600 cells), human pancreas[12–15] (approx. 20,000 cells), and mouse brain[16–19] (approx. 400,000 cells), each with varied complexity.

A unified strategy for cell type annotation

We consider the following general problem setup: Given the access to a set of *n-1* reference studies with labeled cell types, our objective is to assign unknown cells to cell types automatically in a new target study of interest. This is a classical supervised learning problem with the added complexity that, depending on the study, we may expect novel cell types (labels). We address the issue in a manner similar to scmap[2] as follows. First, we define a confidence threshold in our prediction (*thresh = 0.5*). Please note that every cell-type class prediction has a specific confidence level. Subsequently, we assign cells below the threshold with the tag “unknown cell type”, which indicates that the cell has minimal similarity to all the annotated cell types in the reference studies, and is therefore, potentially a novel cell type.

We analyzed five machine-learning models of increasing complexity for our comparison: k-NN classification (knn), logistic regression (lgr) (with both Lasso and ridge penality), Support Vector Machines (svm), Gradient Boosted Trees (xgb), and Multilayer Perceptrons (mlp) (see **Supplementary Methods** for details). To ensure a meaningful and unbiased comparison, we perform a five-fold cross-validated grid search over a wide range of potential parameters for all classifiers. The optimal parameter set for each classifier is selected based on the highest mean accuracy score over the respective five-fold cross-validation test set predictions. We perform the grid

search approach in a leave-one-study-out fashion for all three biological systems, resulting in 14 five-fold cross-validated grid searches for five different classifier types. The classifiers with the final parameter sets are then used to predict the unknown cell types in the target study.

Most machine learning methods are well-suited for cell type annotation given optimal hyper-parameter values

We evaluated cell-type annotation performance of models based on the accuracy of predicting cell types in a holdout study. We particularly measured the accuracy based on receiver operator characteristics per cell type to determine the best performing model across all classifier types (see **Fig. 1a**). Subsequently, we determined the recall of the best performing model and present it as a confusion matrix of actual and predicted cell types in the target study (see **Fig. 1b** and **Supplementary Figure 1**). The final evaluation score is the accuracy of the top performing classifier for predicting all cell-type classes calculated on the target study (see **Fig. 1c**).

For the small-scale embryonic development datasets, we observed considerable differences in performance across models in the same dataset, which could be explained by the small sample size of the training data. For example, the dataset from Deng et al.[9] comprises 314 cells, and the training dataset comprises 284 cells. However, the overall accuracy of the prediction across datasets was not correlated with the sample size ratio of the training and test dataset. Instead, the differences in performance across models on the same test dataset decreased with increasing sample size in the training dataset.

When we used tissue studies of the human pancreas[12–15] and the mouse brain[16–19], all models performed equally well on the prediction task. Smith et al.[20] reported a similar result in bulk RNA-seq data. We achieved a high prediction accuracy and recall in human pancreas studies. We have illustrated the cell-type prediction for the Baron[13] dataset with MLP (see **Fig. 1b**). Notably, MLP could not predict several rare cell types at all (epsilon cells and mast cells), while it predicted abundant cell types accurately. In addition, the reference dataset was more complex compared to the predicted dataset, and the prediction favored rare cell types, such as mesenchymal cells and PSCs, which the authors of the original publication did not observe. Therefore, human expert labels are not necessarily reliable or accurate, particularly in the case of rare cell types.

In case of the large mouse brain datasets, we removed the knn, svm, and *scmap-cell* models due to infeasible runtime and memory usage. The accuracy scores for each classifier varied minimally for every holdout study and were generally comparable across all holdout studies. We even observed a good performance on the 10X Genomics prenatal mouse brain dataset (E18 stage) even though the training data were obtained from postnatal mice. The developing brain undergoes considerable maturation processes before and after birth, which are reflected in the transcriptomes of the respective cells. Here, the lower transcriptional similarity between test data and training data did not result in lower prediction accuracy scores.

Overall, we found that the deep learning models did not outperform classical machine-learning models, such as logistic regression and svm (pairwise wilcoxon rank sum test, see **Supplementary Tables 1-2**), in the prediction/annotation of cell types in new data sets. Instead, factors, such as similarities between training and test data, data complexity in terms of cell types, and mere sample size affect the accuracy of annotation. Nevertheless, model tuning through grid search considerably improved accuracy; in general, all tested classifiers outperformed *scmap-cell*[2] based on accuracy scores (pairwise wilcoxon rank sum test, see **Supplementary Tables 1-2**, see **Fig. 1c**).

Discussion and conclusion

We tested several supervised classifiers in small, medium, and large-scale datasets. We compared optimally tuned classification models with various degrees of complexity using a global hyperparameter search to obtain optimal parameter sets. It is not surprising that the performance of all the trained models improved significantly compared to the model *scmap-cell,* which uses a fixed set of parameters. Clearly, cell-type annotation accuracy benefits from the screening for optimal parameters independent of the model. In addition, the accuracy of the various classifiers is largely similar for each holdout study across numerous data sets, where the differences in the large datasets are lower compared to the differences in small datasets.

Why does deep learning not outperform classical machine-learning methods, at least in the large and complex droplet-based data sets with more than half a million cells? The answer to the question becomes clear when we examine how cell types are manually annotated: we annotate groups of cells, but we apply simple rules such as presence or absence of marker genes. Therefore, manual cell type annotation is a rule-based approach rather than cluster-based. Mathematically, we project the cells in a separable halfspace, where a linear classifier separates two cell types and such a classifier is identical to the last layer of a neural network as well as to the decision boundary of a logistic regression model and a support vector machine. Thus, any machine learning and deep learning model derives linear rules to assign labels to every single cell, in contrast to learning a more complex label.

The term ‘complex labels’ describes class labels, which do not arise from simple single feature classification or Euclidean distance comparisons in the raw data. In contrast to classical machine learning models, deep learning models identify nonlinear representations (or latent spaces), which have little resemblance with the similarity structure of the original data. We illustrated this feature on the *cifar10* dataset (60,000 images in 10 mutually exclusive classes; see **Supplementary Fig. 2**). Here, the neural network recovers a clean separation between the classes (**Supplementary Fig. 2a**), which are not recovered otherwise (**Supplementary Fig. 2b**). In the case of image data, a deep learning model creates a nonlinear connection to a simpler latent space, where linear decision boundaries accurately discriminate between the complex labels (**Supplementary Fig. 2a**). In contrast to image classes, the annotation of cell types in single cell data is a non-complex label, because a clustering of data in the Euclidean

space provides a sufficient grouping of similar cells, *i.e.* the Euclidean distance induces a biologically interpretable structure of the data.

Hence, to learn interesting and generalizable feature spaces, particularly interesting for subsequent transfer learning[21], more complex labels are needed. Such labels might be functional responses to drug treatment, gene perturbation or other stimulations, which may not trivially be visible on expression clusters. For cell lines, retrieving such information requires creative experimental methods such as traceable genetic manipulation[22,23], which allows tracing cell fates over several generations. In tissue samples, the transcriptome of cells depends on its direct neighbourhood[24,25], *i.e.* when we combine spatial information with RNA-seq, we obtain immediately a complex label similar to images. When labels become more complex, we expect the benefits of complex deep learning-based classifiers to increase.

While waiting for the emergence of such labels, the scientific community has developed a series of unsupervised deep-learning approaches, including batch effect correction[26], single-cell perturbation prediction[27], and dropout effects denoising[28]. Deep learning clustering tools such as DCA[28] and scVI[26] help to improve on clustering. For example, DCA and scVI learn a data representation, which is dropout free and batch-effect free, leading to more accurate clusters. However, the prediction of cell types is cell based, not cluster based. Therefore, a better clustering does not address the lack of complexity in the cell type annotation, it reduces noise, increases accuracy and reduces occasional inconsistencies.

In general, the application of neural networks in supervised cell-type annotation provides practical advantages in terms of flexible applicability, highly optimized training libraries, and access to learned feature representation in latent spaces for downstream analysis and visualization[21].

In the present study, we were not able to observe any significant improvement in classification accuracy based on neural networks, in contrast to other fields, such as computer vision. Therefore, the “ImageNet moment” in single-cell genomics is only attainable through more complex labels that complement the genomic information.

Figure legend

Figure 1: All supervised learning models for cell-type annotation has similar performance as measured by accuracy.

a) Receiver operating characteristics for the ductal cell type (Baron[13] data as holdout study) in all tested models (except for *scmap-cell*, which does not return the prediction probability for all classes). The area under curve (AUC) is given in parentheses in the legend. We observe little difference across classifier types.
b) Confusion matrix for the recall of the MLP in pancreas data with Baron[13] data as holdout study.
c) Accuracy scores for knn, lgr, svm, xgb, mlp in each study using the optimal parameter set and *scmap-cell*. In the case of mouse brain data, we had to leave out knn, svm and scmap-cell due to infeasible runtime or memory usage.

References

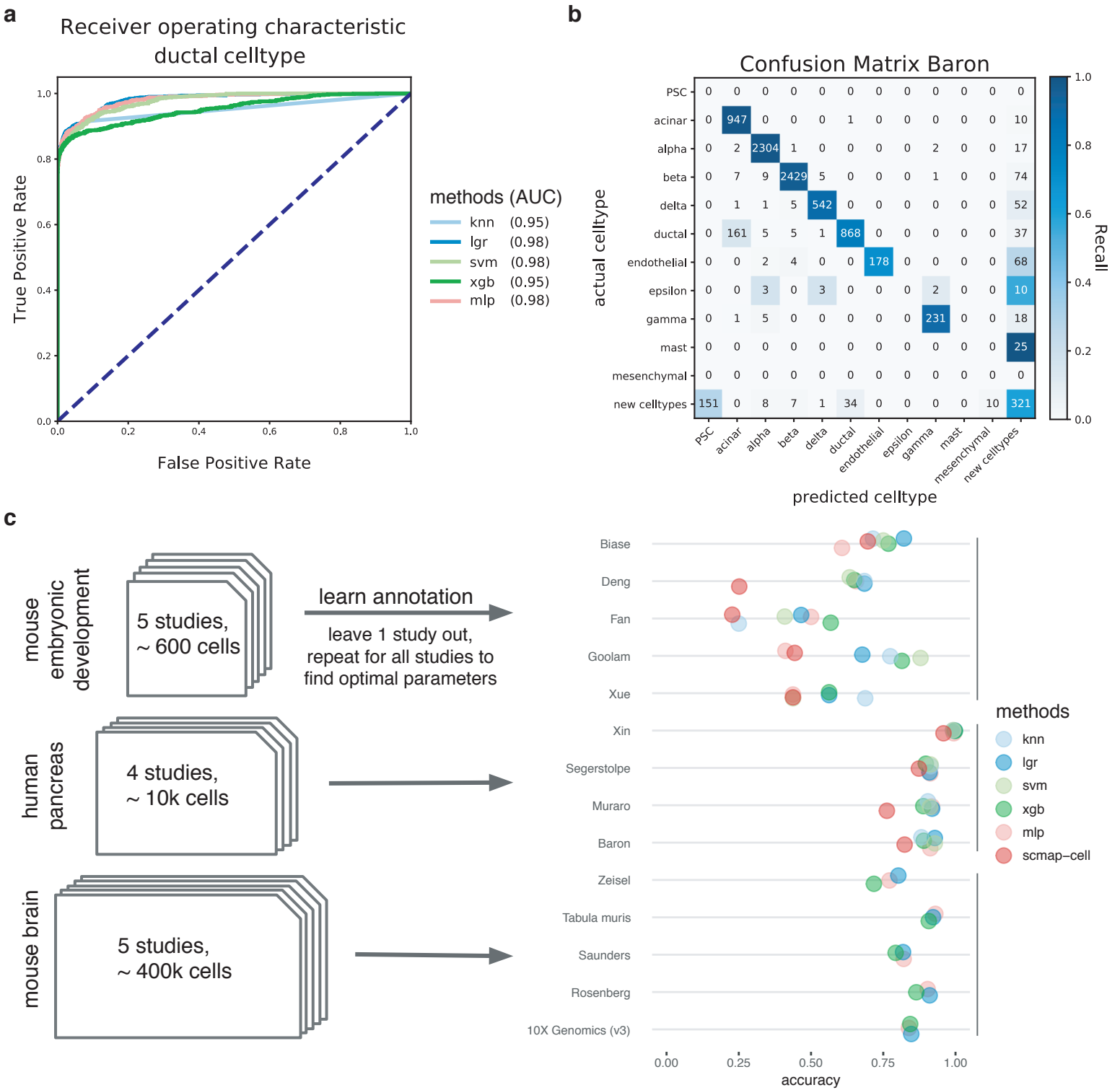
1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).
2. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
3. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
4. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
5. Stuart, T. *et al.* Comprehensive integration of single cell data. *bioRxiv* (2018). doi:10.1101/460147
6. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *bioRxiv* (2019). doi:10.1101/538652
7. Biase, F. H., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24**, 1787–1796 (2014).
8. Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165**, 61–74 (2016).
9. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
10. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
11. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in

mouse preimplantation embryos. *Genome Biol.* **16**, 148 (2015).

12. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).
13. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346–360.e4 (2016).
14. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab.* **24**, 608–615 (2016).
15. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385–394.e3 (2016).
16. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
17. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999–1014.e22 (2018).
18. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
19. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015–1030.e16 (2018).
20. Smith, A. M. *et al.* Deep learning of representations for transcriptomics-based phenotype prediction. *bioRxiv* (2019). doi:10.1101/574723
21. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0122-6
22. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* (2019). doi:10.1038/s41586-019-1184-5
23. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
24. Satija, R., Farrell, J. A., Gennert, D. & Schier, A. F. Spatial reconstruction of single-cell gene expression data. *Nature* (2015).

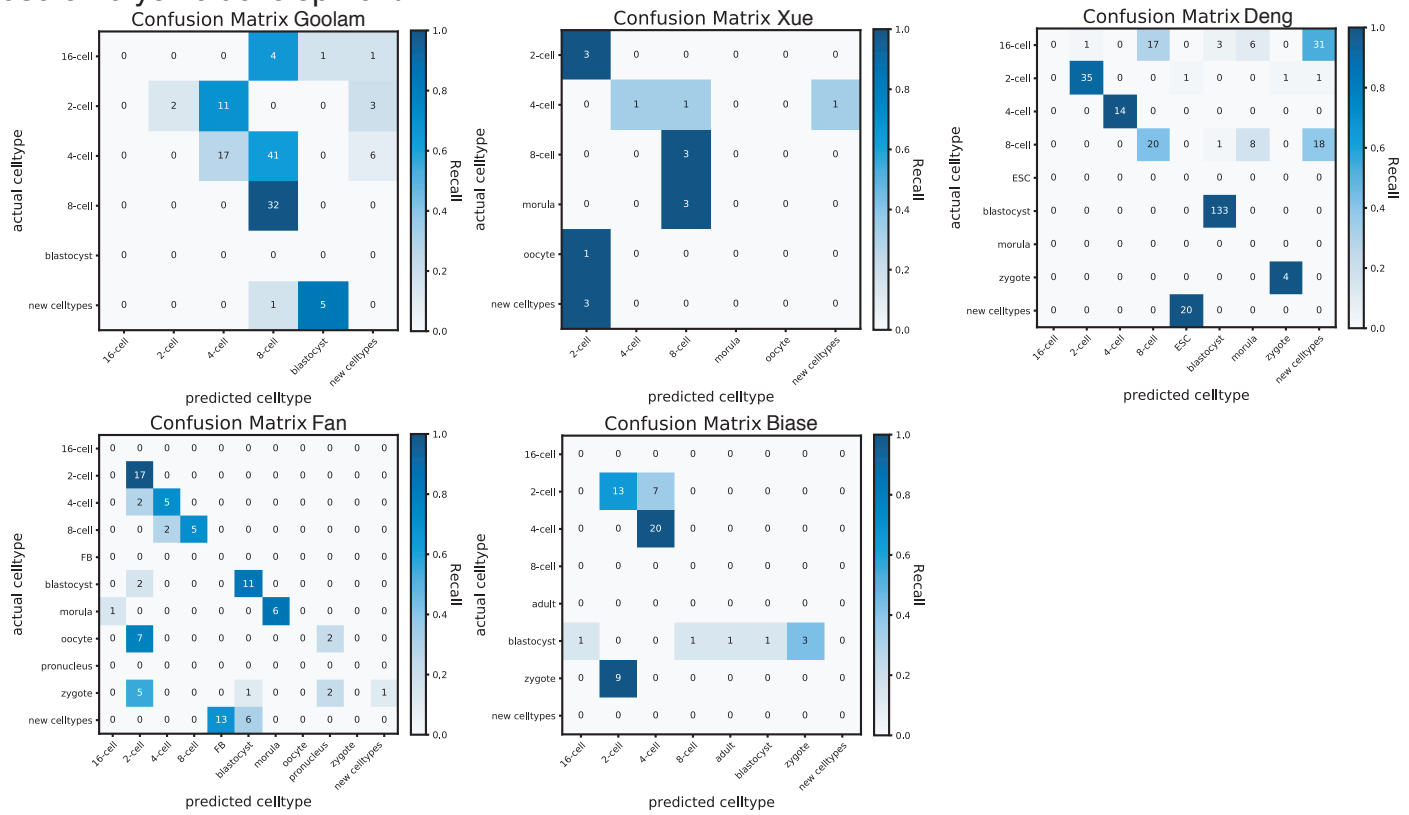
25. Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4231
26. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
27. Lotfollahi, M., Alexander Wolf, F. & Theis, F. J. Generative modeling and latent space arithmetics 4 predict single-cell perturbation response across 5 cell types, studies and species. *bioRxiv* (2018). doi:10.1101/478503
28. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).

Figure 1

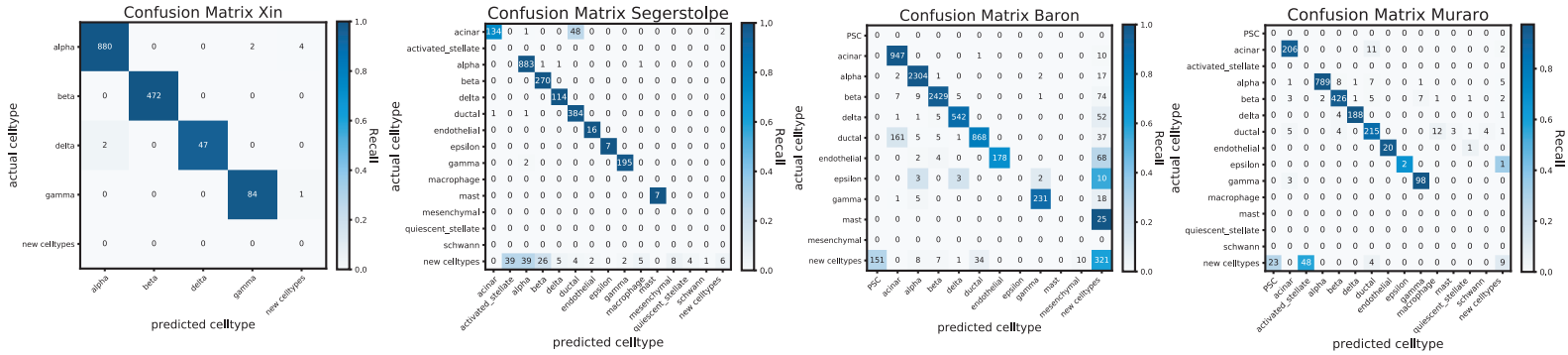

a
Receiver operating characteristic
ductal celltype
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
methods (AUC)
knn (0.95)
lgr (0.98)
svm (0.98)
xgb (0.95)
mlp (0.98)
b
Confusion Matrix Baron
actual celltype
predicted celltype
Recall
PSC 0 0 0 0 0 0 0 0 0 0 0 0
acinar 0 947 0 0 0 1 0 0 0 0 0 10
alpha 0 2 2304 1 0 0 0 0 2 0 0 17
beta 0 7 9 2429 5 0 0 0 1 0 0 74
delta 0 1 1 5 542 0 0 0 0 0 0 52
ductal 0 161 5 5 1 868 0 0 0 0 0 37
endothelial 0 0 2 4 0 0 178 0 0 0 0 68
epsilon 0 0 3 0 3 0 0 0 2 0 0 10
gamma 0 1 5 0 0 0 0 0 231 0 0 18
mast 0 0 0 0 0 0 0 0 0 0 0 25
mesenchymal 0 0 0 0 0 0 0 0 0 0 0 0
new celltypes 151 0 8 7 1 34 0 0 0 0 10 321
PSC acinar alpha beta delta ductal endothelial epsilon gamma mast mesenchymal new celltypes
c
mouse embryonic development
5 studies, ~ 600 cells
learn annotation
leave 1 study out, repeat for all studies to find optimal parameters
human pancreas
4 studies, ~ 10k cells
mouse brain
5 studies, ~ 400k cells
Biase
Deng
Fan
Goolam
Xue
Xin
Segerstolpe
Muraro
Baron
Zeisel
Tabula muris
Saunders
Rosenberg
10X Genomics (v3)
methods
knn
lgr
svm
xgb
mlp
scmap−cell
0.00
0.25
0.50
0.75
1.00
accuracy

Supplementary Figure 1: Confusion matrices for all tested datasets

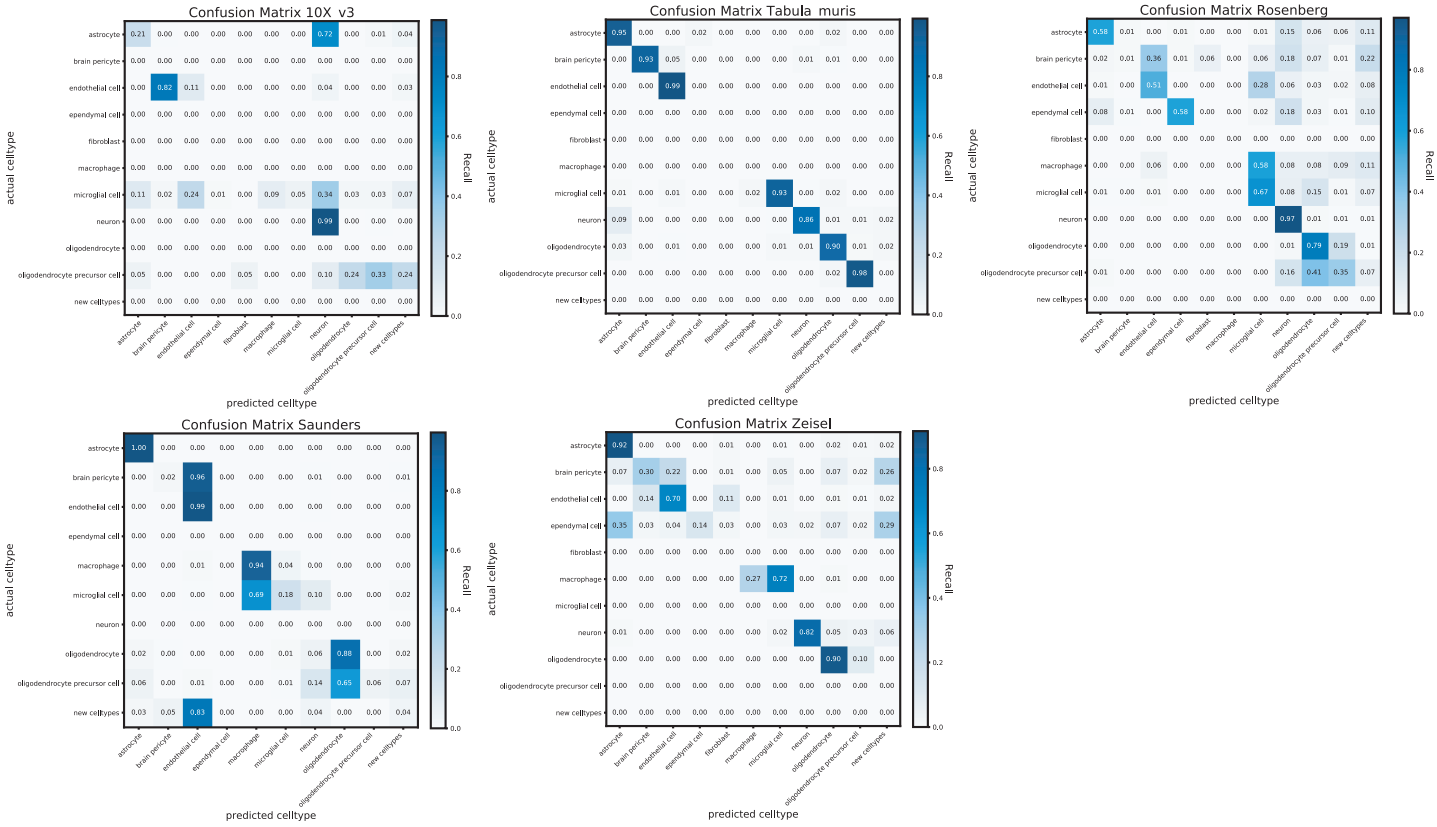
a Mouse embryonic development



b Human pancreas



c Mouse brain



Supplementary Figure 2: t-SNE plots of the CIFAR-10 data set, colored by the 10 classes

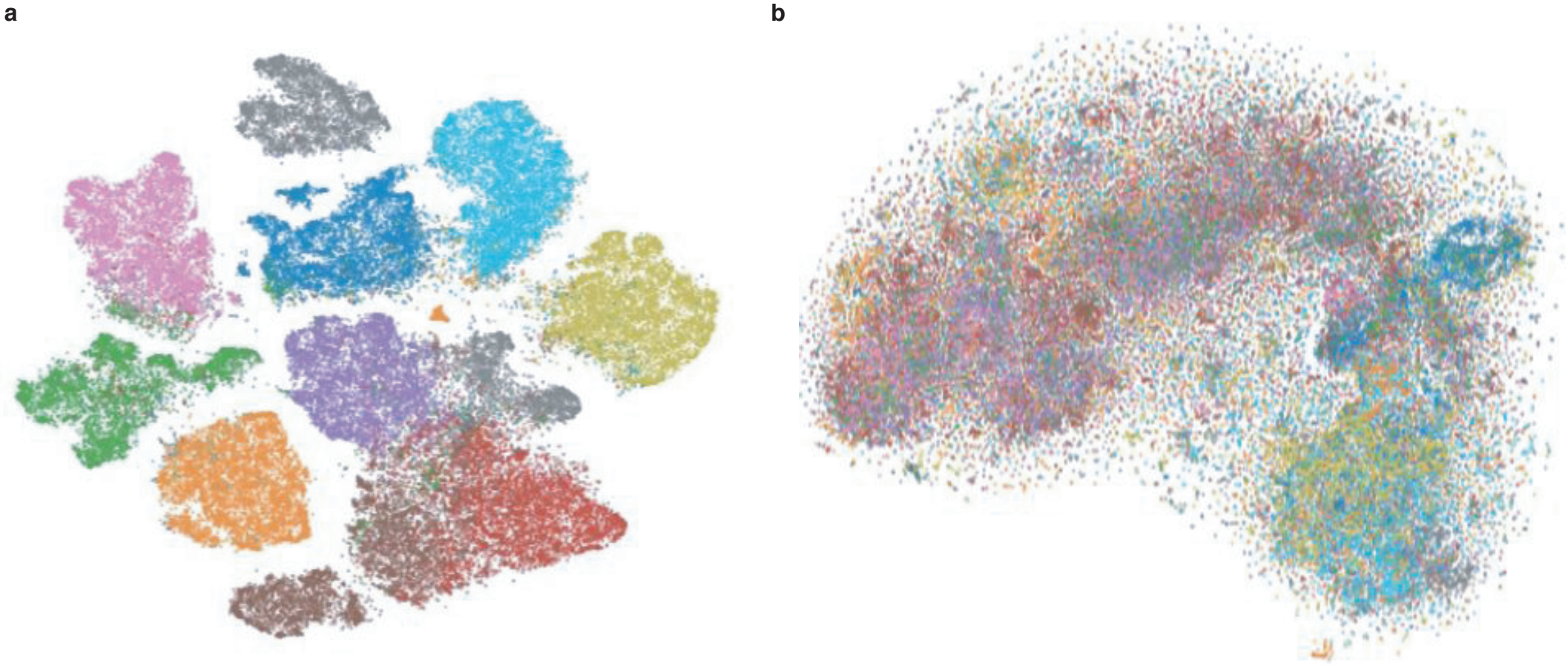


Figure legends

Supplementary Figure 1: Confusion matrices for the recall of the MLP for every data set as holdout study.

a. Confusion matrices for mouse embryonic development studies.
b. Confusion matrices for human pancreas studies.
c. Confusion matrices for mouse brain studies.

Supplementary Figure 2: t-SNE plots of the CIFAR-10 data set, colored by the 10 classes.

a. t-SNE computed on the learned features of a neural network.
b. t-SNE computed on the raw input images.

Supplementary Methods

Cell-type annotation algorithms

Logistic regression (lgr): We used the sklearn (0.19.1) implementation for logistic regression in python 3.6.8 with the following parameters:

Parameter	**mouse embryo**	**human pancreas**	**mouse brain**
C	1	0.01	1
class_weight	None	None	None
dual	False	False	False
fit_intercept	True	True	True
intercept_scaling	1	1	1
max_iter	100	100	100
multi_class	ovr	ovr	ovr
n_jobs	1	1	1
penalty	l1	l2	l1
random_state	42	42	42
solver	liblinear	liblinear	liblinear
tol	0.0001	0.0001	0.0001
warm_start	False	False	False

k-NN classifier (knn): We used the sklearn (0.19.1) implementation for the k-NN classifier in python 3.6.8 with the following parameters:

Parameter	mouse embryo	human pancreas	mouse brain
algorithm	auto	auto	-
leaf_size	30	30	-
metric	minkowski	minkowski	-
metric_params	None	None	-
n_neighbors	5	30	-
p	2	2	-
weights	distance	distance	-

Support vector machine (svm): We used the sklearn (0.19.1) implementation for the support vector machine classifier in python 3.6.8 with the following parameters:

Parameter	mouse embryo	human pancreas	mouse brain
C	0.1	0.1	-
cache_size	200	200	-
class_weight	None	None	-
coef0	0.0	0.0	-
decision_function_shape	ovr	ovr	-
degree	3	3	-
gamma	auto	auto	-
kernel	linear	linear	-
probability	True	True	-
shrinking	True	True	-
tol	0.001	0.001	-
random_state	42	42	-

Gradient boosted trees (xgb package): We used the xgboost (0.72) implementation for the gradient boosted trees classifier in python 3.6.8 with the following parameters:

Parameter	mouse embryo	human pancreas	mouse brain

base_score	0.5	0.5	0.5
booster	gbtree	gbtree	gbtree
colsample_bylevel	1	1	1
colsample_bytree	0.75	0.75	1.0
gamma	0	0	0
learning_rate	0.1	0.1	0.1
max_detla_step	0	0	0
max_depth	4	4	4
min_child_weight	1	1	1
missing	nan	nan	nan
n_estimators	50	100	500
reg_alpha	0	0	0
reg_lambda	1	1	1
scale_pos_weigth	1	1	1
subsample	0.75	0.75	1.0
tree_method	hist	hist	hist

Multilayer-perceptron (mlp): We used the sklearn (0.19.1) implementation for the multilayer-perceptron classifier in python 3.6.8 with the following parameters:

Parameters	**embryo**	**human pancreas**	**mouse**
activation	relu	relu	relu
alpha	0.0001	0.0001	0.0001
batch_size	auto	auto	auto
beta_1	0.9	0.9	0.9
beta_2	0.999	0.999	0.999
early_stopping	True	True	True
epsilon	1e−08	1e−08	1e−08
hidden_layer_sizes	(100,)	(100,100)	(100,)
learning_rate	invscaling	invscaling	invscaling

learning_rate_init	0.01	0.001	0.001
max_iter	200	200	200
momentum	0.9	0.9	0.9
nesterovs_momentum	True	True	True
power_t	0.5	0.5	0.5
solver	adam	adam	adam
tol	0.001	0.001	0.001
validation_fraction	0.1	0.1	0.1
warm_start	False	False	False

Scmap-cell: We used *scmap-cell* as implemented in scmap (v. 1.4.1) in R (v. 3.5.2) based on the accompanying vignette with default parameters.

Parameter Optimization via grid search

We used the sklearn (0.19.1) implementation of GridSearchCV in python 3.6.8 to perform the five-fold cross-validated grid search for all of the models in all the three biological systems. For each classifier we defined a parameter grid.

XGBoost

Parameter	**Range**
max_depth	[2,4,8,10]
n_estimators	[20,50,100,500]
subsample	[0.75,1.00]
colsample_bytree	[0.75,1.00]

Logistic Regression

Parameter	**Range**
C	[0.01,0.1,1,10]
penalty	[“l1,”“l2”]

Multi-Layer-Perceptron

Parameter	**Range**

hidden_layer_sizes	[(100), (100,100), (500), (500,500), (1000,1000,1000)]
learning_rate_init	[0.01,0.001]
activation	[“relu’,”sigmoid”]

Support Vector Machine

Parameter	**Range**
C	[0.01,0.1,1,10]
kernel	[”poly”,”linear”]

kNN-Classifier

Parameter	**Range**
n_neighbors	[5,10,30]
weights	[”uniform”,”distance”]

Data availability and pre-processing

Mouse embryonic development studies:
We used 5 different SMARTseq-based scRNA-seq studies[7–11], consisting of 56, 314, 88, 124 and 16 cells, respectively, with the following accession ids: E-GEOD-57249, E-MTAB-3321, GSE53386, E-GEOD-45719 and E-GEOD-44183. Fastq files were mapped to Ensembl[29] mouse transcriptome (version GRCm38.p5.87) with Salmon[30] (version 0.8.2, kmer = 21 to tolerate different read length). We tested the cell type annotation models on log-normalized data without further gene filtering or quality control.

Human Pancreas studies:
We used 4 different human pancreas studies[12–15], consisting of 3514 cells (Smart-Seq2), 8569 cells (inDrop), 1600 cells (SMARTer) and 2126 cells (CEL-Seq2), respectively, with the following accession ids: E-MTAB-5061, GSE84133, GSE81608 and GSE85241. Please note that a pre-processed and annotated version of the count matrices is provided by the Hemberg lab (https://hemberg-lab.github.io/scRNA.seq.datasets/human/pancreas/), which we used as resource. We tested the cell type annotation models on log-normalized data without further gene filtering or quality control.

Mouse brain studies:
We used 4 different scRNA-seq studies[16–19] and a single cell gene expression dataset for 10k mouse brain cells (E18) (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/neuron_10k_v3) from 10X Genomics, consisting of 133,180 cells (SPLiT-seq protocol), 160,796 cells (10X Genomics Chromium), 7,856 myeloid and non-myeloid cells (SMART-Seq2 protocol), 109,799 cells (Drop-seq protocol), and 11,741 cells (10X Genomics Chromium), respectively.

Rosenberg et al. have deposited the raw count matrix on GEO with accession ID GSE110823. Zeisel et al. have deposited the data as web resource (http://mousebrain.org/), from which we downloaded the annotated count matrix as loom file 'L5_all.loom'. FAC-sorted mouse brain tissue data (myeloid and non-myeloid cells) from Tabula Muris are available through https://tabula-muris.ds.czbiohub.org/ and figshare (https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733). Saunders et al. have deposited the data as web resource (http://dropviz.org/), from which we downloaded count matrices per cell type in DGE file format ('DGE by Class' section). A single cell gene expression dataset for 10k mouse brain cells (E18) (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/neuron_10k_v3) was downloaded from 10X Genomics.
All datasets except for the 10X Genomics 10k mouse brain dataset were pre-filtered and annotated. We unified any differences in notation across the datasets and log-normalized the count matrices. We used the scanpy[31] framework (v. 1.4+14.gd4a7a2d) to carry out pre-processing. We filtered the 10X Genomics 10k mouse brain dataset by removing all cells with less than 650 counts and less than 500 expressed genes. All genes detected in less than 2 counts were removed. The final count matrix was log-normalized. Then, data were clustered using louvain clustering with resolution 1.0. Further, we annotated the clusters through marker gene expression.

Cifar10 image data set:
60,000 images annotated in 10 mutually exclusive classes are available under https://www.cs.toronto.edu/~kriz/cifar.html.

Significance test for overall performance

We performed a paired, one-sided, pairwise wilcoxon rank sum test in R (v. 3.5.2) on all studies (n = 9 for pairwise comparisons of knn, svm, lgr, xgb, mlp, and scmap-cell, and n = 14 for pairwise comparisons of lgr, xgb, and mlp) to assess the overall performance of the classifiers with Benjamini-Hochberg correction for multiple testing (function pairwise.wilcox.test).

Supplementary Table 1: p-values of the paired, one-sided, and pairwise wilcoxon rank sum test with the alternative hypothesis "Method in row performed with less accuracy than method in column." Data: Mouse embryonic development and human pancreas studies (n = 9).

	knn	lgr	mlp	svm	xgb
lgr	0.8433				
mlp	0.5565	0.1342			
svm	0.8433	0.2176	0.7961		
xgb	0.7961	0.6947	0.8496	0.7961	
scmap-cell	0.0098	0.0098	0.1342	0.0267	0.0098

Supplementary Table 2: p-values of the paired, one-sided, and pairwise wilcoxon rank sum test with the alternative hypothesis "Method in row performed with less accuracy than method

in column." Data: Mouse embryonic development, human pancreas studies and mouse brain studies (n = 14).

	lgr	mlp
mlp	0.087	
xgb	0.138	0.524

References

29. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).
30. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017).
31. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).